

175
NOVEMBER 1969

A HEURISTIC PROGRAM FOR SOLVING
A SCIENTIFIC INFERENCE PROBLEM:
SUMMARY OF MOTIVATION AND IMPLEMENTATION*

by

Joshua Lederberg, Georgia L. Sutherland,
Bruce G. Buchanan, Edward A. Feigenbaum**

ABSTRACT: The primary motivation of the Heuristic DENDRAL project is to study and model processes of inductive inference in science, in particular, the formation of hypotheses which best explain given sets of empirical data. The task chosen for detailed study is organic molecular structure determination using mass spectral data and other associated spectra. This paper first summarizes the motivation and general outline of the approach. Next, a sketch is given of how the program works and how good its performance is at this stage. The paper concludes with a comprehensive list of publications of the project.

*This research was supported mainly by the Advanced Research Projects Agency and also by the National Aeronautics and Space Administration. Portions of this paper have appeared in a summary report to NASA, and in Machine Intelligence 5 (Edinburgh University Press).

**We gratefully acknowledge the collaboration of Mr. Allan Delfino, Dr. Alan Duffield, Dr. Gustov Schroll, and Professor Carl Djerassi.

A HEURISTIC PROGRAM FOR SOLVING
A SCIENTIFIC INFERENCE PROBLEM:
SUMMARY OF MOTIVATION AND IMPLEMENTATION*

Joshua Lederberg, Georgia L. Sutherland

Bruce G. Buchanan, Edward A. Feigenbaum**

Part I: Motivation

The "scientific method" involves two very different kinds of intelligent behavior, sometimes called induction and deduction respectively. A theory is somehow "induced", sometimes out of sheer speculation, in order to account for some hitherto baffling or provocative observations of nature. Then, the theory is applied deductively, i.e., logically or mathematically rigorous conclusions are made. If the theory is true, then certain results must be obtained. Philosophers of science are now generally agreed that a theory can never be proven or logically derived from factual data. We accept a theory as true when it has made some new predictions, different from the predictions of other theories, which survive the test of experimental measurement.

The process of logical deduction follows rules which, at least at an elementary level, are well understood. Correspondingly, computers have been extensively used for deductive calculations; for example, to predict the path of a ballistic projectile in the gravitational fields of the solar system. When discrepancies are found, the theory must be questioned at some level; for example, mascons in the moon are postulated as a simpler explanation of certain perturbations than a revision of the laws of gravitational attraction.

Scientific induction remains a mysterious process, connected to the most "creative" aspects of human thinking, and is difficult to implement on a computer.

The DENDRAL project aims at emulating in a computer program the inductive behavior of the scientist in an important but sharply limited area of science, organic chemistry. Most of our work is addressed to the following problem: Given the data of the mass spectrum of an unknown compound, induce a workable number of plausible solutions, that is, a small list of candidate molecular structures. In order to complete the task, the DENDRAL program then deduces the mass spectrum predicted by the theory of mass spectrometry for each of the candidates, and selects the

most productive hypothesis, i.e., the structure whose predicted spectrum most closely matches the data.

We have designed, engineered, and demonstrated a computer program that manifests many aspects of human problem-solving techniques. It also works faster than human intelligence in solving problems chosen from an appropriately limited domain of types of compounds, as illustrated in the cited publications. (2,3)

Some of the essential features of the DENDRAL program include:

- 1) Conceptualizing organic chemistry in terms of topological graph theory, i.e., a general theory of ways of combining atoms.

- 2) Embodying this approach in an exhaustive HYPOTHESIS GENERATOR. This is a program which is capable, in principle, of "imagining" every conceivable molecular structure.

- 3) Organizing the GENERATOR so that it avoids duplication and irrelevancy, and moves from structure to structure in an orderly and predictable way.

The key concept is that induction becomes a process of efficient selection from the domain of all possible structures. Heuristic search and evaluation is used to implement this "efficient selection". Most of the ingenuity in the program is devoted to heuristic modifications of the GENERATOR. Some of these modifications result in early pruning of unproductive or implausible branches of the search tree. Other modifications require that the program consult the data for cues (pattern analysis) that can be used by the GENERATOR as a plan for a more effective order of priorities during hypothesis generation. The program incorporates a memory of solved sub-problems that can be consulted to look up a result rather than compute it over and over again. The program is aimed at facilitating the entry of new ideas by the chemist when discrepancies are perceived between the actual functioning of the program and his expectation of it.

The attached references report the practical application of DENDRAL as an aid in solving problems of chemical structure. While our main interest in DENDRAL is as a prototype of scientific induction, it may have specific application in guiding the closed-loop automation of an analytical mass spectrometer or general chemical fractionation system.

Part II: Implementation

For this discussion it is sufficient to say that a mass spectrometer is an instrument into which is put a minute sample of some chemical compound and out of which comes data usually represented as a two-dimensional histogram. This is what is referred to here as the mass spectrum. The instrument itself bombards molecules of the compound with electrons, thereby producing ions of different masses in varying proportions. The points on the abscissa of the spectrum represent the masses of ions produced and the points on the ordinate represent the relative abundances of ions of these masses.

The HEURISTIC DENDRAL process of analyzing a mass spectrum consists of three phases. The first, preliminary inference (or planning), obtains clues from the data as to which classes of chemical compounds are suggested or forbidden by the data. The second phase, structure generation, enumerates chemically plausible structural hypotheses which are compatible with the inferences made in phase one. The third phase, prediction and testing (or hypothesis validation), predicts consequences from each structural hypothesis and compares this prediction with the original spectrum to choose the hypothesis which best

explains the data. Corresponding to these three phases are three sub-programs. The program(s) have been described in previous publications, primarily in the book Machine Intelligence 4 (5) and in a series of Stanford Artificial Intelligence Project Memos (4, 5, 7, 11).

The PRELIMINARY INFERENCE MAKER program contains a list of names of structural fragments, each of which has special characteristics with respect to its activity in a mass spectrometer. These are called "functional groups". Each functional group has associated with it a set of spectral values and relationships among these values that are, to the best of our present knowledge, "diagnostic" for the chemical functional group. Other properties of the functional group indicate which other groups are related to this one - as special or general cases.

The program progresses through the group list, checking the conditions for each group. Two lists are constructed for output: GOODLIST enumerates functional groups which might be present, and BADLIST lists functional groups which cannot be in the substance that was introduced to the mass spectrometer.

GOODLIST and BADLIST are the inputs to the STRUCTURE GENERATOR, which is a generator of isomers (topologically

possible graphs) of a given empirical formula (collection of atoms). GOODLIST and BADLIST control and constrain the generation of paths in this space. Each GOODLIST item is treated as a "super atom", so that any functional group inferred from the data by the PRELIMINARY INFERENCE MAKER will be guaranteed to appear in the list of candidate hypotheses output by the STRUCTURE GENERATOR.

The STRUCTURE GENERATOR's operation is based on the DENDRAL algorithm for classifying and comparing acyclic structures.(9) The algorithm guarantees a complete, non-redundant list of isomers of an empirical formula.

The third sub-program is the Mass Spectrum PREDICTOR, which contains what has been referred to as the "complex theory of mass spectrometry". This is a deductive model of the processes which affect a structure when it is placed in a mass spectrometer. Some of these rules determine the likelihood that individual bonds will break, given the total environment of the bond. Other rules are concerned with larger fragments of a structure - like the functional groups which are the basis of the PRELIMINARY INFERENCE MAKER. All these rules are applied (recursively) to each structural hypothesis coming from the STRUCTURE GENERATOR. The result is a list of mass-intensity number pairs, which is the predicted mass spectrum for each candidate molecule.

Any structure is discarded which appears to be inconsistent with the original data (i.e., its predicted spectrum is incompatible with the given spectrum). The remaining structures are ranked from most to least plausible on the basis of how well their spectra compare with the data. The top ranked structure is considered to be the "best explanation".

Thanks to the collaboration of Dr. Gustav Schroll, an NMR (Nuclear Magnetic Resonance) PREDICTOR and INFERENCE MAKER have been added to the program. Thus the program can confirm and rank candidate structures through predictions independent of mass spectroscopy, bringing the whole process more in line with standard accounts of "the scientific method". Thus the HEURISTIC DENDRAL program is expanding from the "automatic mass spectroscopist" to the "automatic analytical chemist". Other analytical tools, such as infra-red spectroscopy, will be incorporated eventually. Only the clumsiness of the language hinders further extensions to conventional "wet chemistry" reactions.

Interaction and interdependence of the three sub-programs of HEURISTIC DENDRAL must be mentioned when discussing these computer programs. Because of the size of the combined programs, it is more practical to run them

separately than to run them together. One supervisor takes care of the interaction by having each sub-program write an output file which is then the input file for the next phase of program operation. The PRELIMINARY INFERENCE MAKER writes the file containing the empirical formula and the GOODLIST and BADLIST to be used by the STRUCTURE GENERATOR. That program, in turn, reads this file, and writes another file containing the single output list of structures which it generates according to the GOODLIST and BADLIST specifications. The PREDICTOR then reads this file to obtain its input, and calculates a mass spectrum for each structure in the file. If other tests such as NMR prediction are to be made on the candidate structures, the supervisor interfaces the appropriate program to these others in the same way.

Three papers have appeared in the Journal of the American Chemical Society (1, 2, 3). The first paper describes the HEURISTIC DENDRAL program and tabulates numbers of chemically plausible isomers for many compounds. This is of particular interest to chemists because it indicates the size of the search space in which structures must be found to match specific data. The second paper explains the application of the program to ketones: the subclass of molecular structures containing the keto

radical. The whole process from preliminary inference (planning) through structure generation and prediction of theoretical spectra was applied to many examples of ketone spectra. The results, in terms of actual structures identified, were encouraging. The third paper explains the application of the program to ethers. Introducing the NMR PREDICTOR contributed to the successful results which are described in the ether paper.

A measure of the program's performance level is provided by comparing the program with professionals. In July (1969) Professor Carl Djerassi, an eminent mass spectroscopist, asked the members of his graduate mass spectrometry seminar to interpret three mass spectra, giving them only the empirical formulas of the structures and stating the fact that they were acyclic structures - just the information given to the program. On the first problem, the program and one graduate student got the correct structure; another graduate student and a post-doctoral fellow were both close, but not correct. On the second problem, the program got the correct answer; two graduate students included the correct answer in undifferentiated sets of two and four structures; while the post-doctoral fellow missed the answer. On the last problem, the program missed the correct structure and the post-doctoral fellow

included it in a pair of equally likely structures. The computer spent approximately two to five minutes on each problem; the chemists spent between fifteen and forty minutes on each. From this small experiment and their own observations, (admittedly sympathetic) mass spectroscopists have said the program performs as well as graduate students and post-doctoral fellows in its limited task domain.

Part III: Commentary

One reason for the high level of performance of the program is the large amount of mass spectrometry knowledge which chemists have imparted to the program. Obtaining this has been one of the biggest bottlenecks in developing the program. It should be understood that there presently is no axiomatic or even well organized theory of mass spectrometry which we could transfer to the program from a text book or from an expert. Most of the chemical theory has been put into the program by a programmer who is not a chemist but who spent many hours in eliciting the theory from the chemist-expert. In many cases the chemist's theory was only tentative or incompletely formulated, so that many iterations of rule formulating, programming, and testing were necessary to bring the DENDRAL program to its present

level of competence.

A few general points of strategy have emerged from the DENDRAL effort. With regard to the theoretical knowledge of the task domain in the program, we believe that the following considerations are important.

1) It is important that the program's "theory of the real world" be centralized and unified. Otherwise, during the evolution of this theory, the program will inevitably accumulate inconsistencies - as when one module of the theory expects organic compounds to contain sulfur, while sulfur is denied in another portion of the theory.

2) It would be advantageous for the program to derive planning (Preliminary Inference) cues from its own theory, by introspection, rather than from external data which may not yet have been assimilated into its theory. The success of the program depends in every case on the validity of the theory, so there is no use going beyond it. It is more efficient for the computer to generate hypothetical spectra and search for the relevant "diagnostic" patterns in them than to wait for experimental data. The theory should be responsive to the data; then the list of inference cues should be generated from the theory.

3) Separating the theory from the routines which use it facilitates changing the theory to improve it, on the one hand, or to experiment with variations of it, on the other. Although scattering the theory in the program's LISP code increases running efficiency, it seems more desirable, at this point, to increase the program's flexibility. This has led us to design the programs in a form we refer to as "table-driven". Reference 11 contains a more complete discussion of this effort.

BIBLIOGRAPHY

- (1) J. Lederberg, G. L. Sutherland, B. G. Buchanan, E. A. Feigenbaum, A. V. Robertson, A. M. Duffield, and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference I. The Number of Possible Organic Compounds: Acyclic Structures Containing C, H, O and N". *Journal of the American Chemical Society*, 91:11 (May 21, 1969).
- (2) A. M. Duffield, A. V. Robertson, C. Djerassi, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, and J. Lederberg, "Application of Artificial Intelligence for Chemical Inference II. Interpretation of Low Resolution Mass Spectra of Ketones". *Journal of the American Chemical Society*, 91:11 (May 21, 1969).
- (3) G. Schroll, A. M. Duffield, C. Djerassi, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, and J. Lederberg, "Application of Artificial Intelligence for Chemical Inference III. Aliphatic Ethers Diagnosed by Their Low Resolution Mass Spectra and NMR Data". *Journal of the American Chemical Society* (in press).
- (4) G. Sutherland, "HEURISTIC DENDRAL: A Family of LISP Programs", to appear in D. Bobrow (ed), *LISP Applications* (also Stanford Artificial Intelligence Project Memo No. 80).
- (5) B. G. Buchanan, G. L. Sutherland, and E. A. Feigenbaum, "HEURISTIC DENDRAL: A Program for Generating Explanatory Hypotheses in Organic Chemistry". In *Machine Intelligence 4* (B. Meltzer and D. Michie, eds) Edinburgh University Press (1969), (also Stanford Artificial Intelligence Project Memo No. 62).
- (6) C. W. Churchman and B. G. Buchanan, "On the Design of Inductive Systems: Some Philosophical Problems". *British Journal for the Philosophy of Science*, (Autumn 1969).
- (7) J. Lederberg and E. A. Feigenbaum, "Mechanization of Inductive Inference in Organic Chemistry", in B. Kleinmuntz (ed) *Formal Representations for Human Judgment*, (Wiley, 1968) (also Stanford Artificial Intelligence Project Memo No. 54).
- (8) E. A. Feigenbaum, "Artificial Intelligence: Themes in the Second Decade". *Proceedings of the IFIP68 International Congress*, Edinburgh, August 1968 (in press), (also Stanford Artificial Intelligence Project Memo No. 67).

(9) J. Lederberg, "DENDRAL-64 - A System for Computer Construction, Enumeration and Notation of Organic Molecules as Tree Structures and Cyclic Graphs", (technical reports to NASA, also available from the author and summarized in (14)).

(9a) Part I. Notational algorithm for tree structures (1964) CR.57029

(9b) Part II. Topology of cyclic graphs (1965) CR.68898

(9c) Part III. Complete chemical graphs; embedding rings in trees (1969)

(10) D. A. Waterman, "Machine Learning of Heuristics", PhD Dissertation (Stanford University Computer Science Department), (also Stanford Artificial Intelligence Project Memo No. 74).

(11) B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, "Toward an Understanding of Information Processes of Scientific Inference in the Context of Organic Chemistry", in Machine Intelligence 5, (B. Meltzer and D. Michie, eds) Edinburgh University Press (in press), (also Stanford Artificial Intelligence Project Memo no. 99).

(12) J. Lederberg, "Systematics of organic molecules, graph topology and Hamilton circuits. A general outline of the DENDRAL system." NASA CR-48899 (1965)

(13) J. Lederberg, "Online computation of molecular formulas from mass number." NASA CR-94977 (1968)

(14) J. Lederberg, "Topology of Molecules", in The Mathematical Sciences - A Collection of Essays, (ed.) Committee on Support of Research in the Mathematical Sciences (COSRIMS), National Academy of Sciences - National Research Council, M.I.T. Press, (1969), pp. 37-51.

(15) J. Lederberg, "Computation of Molecular Formulas for Mass Spectrometry", Holden-Day, Inc. (1964).